Shadow Manifold Hamiltonian Monte Carlo

Chris van der Heide[1] UQ
Liam Hodgkinson UC Berkeley, ICSI
Fred Roosta UQ, ICSI
Dirk P. Kroese UQ

Abstract

Hamiltonian Monte Carlo and its descendants have found success in machine learning and computational statistics due to their ability to draw samples in high dimensions with greater efficiency than classical MCMC. One of these derivatives, Riemannian manifold Hamiltonian Monte Carlo (RMHMC), better adapts the sampler to the geometry of the target density, allowing for improved performances in sampling problems with complex geometric features. Other approaches have boosted acceptance rates by sampling from an integrator-dependent "shadow density" and compensating for the induced bias via importance sampling. We combine the benefits of RMHMC with those attained by sampling from the shadow density, by deriving the shadow Hamiltonian corresponding to the generalized leapfrog integrator used in RMHMC. This leads to a new algorithm, shadow manifold Hamiltonian Monte Carlo, that shows improved performance over RMHMC, and leaves the target density invariant.

1 Introduction

Hamiltonian Monte Carlo (HMC) was first introduced by Duane et al. (1987) in the context of simulating lattice field theories in quantum chromodynamics. It gained further mainstream success following its introduction to the machine learning and computational statistics communities for training Bayesian neural networks (Neal, 1996). It has since become an indispensable tool for Bayesian inference at large, finding a plethora of applications in scientific and industrial fields.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

The key advantage of HMC over many of its competitors is its ability to draw samples that are large distances apart by evolving them via Hamiltonian dynamics. The acceptance rate depends on the error accumulated along the sample trajectory, and remains large even in moderately high dimensions. This is in no small part due to its deep connection with geometry and physics, which inform its rich theoretical foundations (Betancourt et al., 2017; Barp et al., 2018). However, HMC's performance deteriorates if either the step size or the size of the system becomes too large, or the target density is poorly behaved. More numerical error leads to lower sample acceptance, which induces heavy autocorrelation, necessitating a larger sample size and thus higher computational costs.

One approach to ease this burden is to exploit the structure of the numerical integrator error and instead target the density corresponding to a modified, or *shadow*, Hamiltonian. This leads to a higher sample acceptance rate, shown in Figure 1, at the cost of some induced bias. This bias is easily quantified, and importance sampling is then performed to compensate (Radivojević & Akhmatskaya, 2019). Due to the structure of the shadow Hamiltonian, *momentum refreshment* requires another Metropolis–Hastings step to ensure sampler consistency. This results in a non-reversible sampler, and allows for partial momentum retention (Horowitz, 1991; Kennedy & Pendleton, 2001; Akhmatskaya & Reich, 2006; Sohl-Dickstein, 2012; Sohl-Dickstein et al., 2014), which may be of interest on its own.

In their landmark paper, Girolami & Calderhead (2011) proposed *Riemannian manifold Hamiltonian Monte Carlo* (RMHMC). This algorithm draws upon ideas from information geometry (Amari, 2016) to generalize HMC by traversing a Riemannian manifold. While other special classes of Riemannian manifolds have recently been studied (Barp et al., 2018, 2019), the original implementations target Bayesian posterior densities, and excel in the presence of complex geometric features. However, to date, there have been no efforts to develop a shadow Hamiltonian corresponding to the generalized leapfrog algorithm used in RMHMC.

[1]chris.vdh@gmail.com

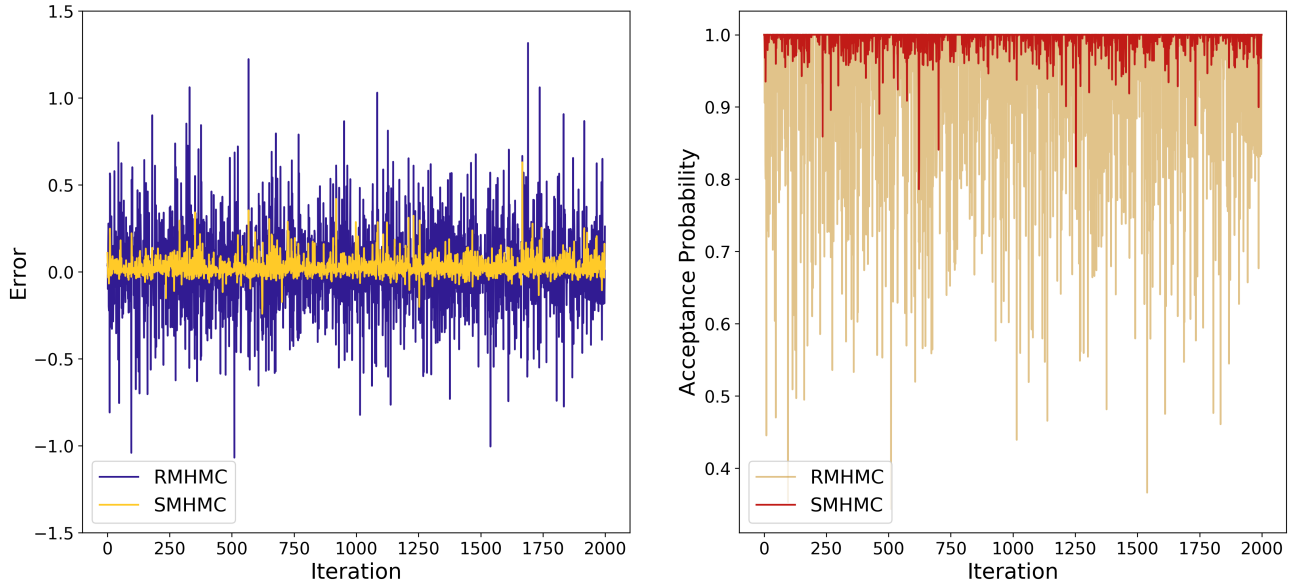

Figure 1: Error due to Hamiltonian drift (left), and corresponding Metropolis–Hastings acceptance probabilities (right), in posterior samples from a Bayesian logistic regression model using the Australian credit dataset. 2000 proposed parameter samples were produced via RMHMC and our proposed SMHMC with $h = 0.5, L = 6$.

Contributions. Our contributions in this work can be summarized as follows.

I. A fourth-order shadow Hamiltonian that is preserved by the generalized leapfrog integrator up to $\mathcal{O}(h^4)$ in the step size h is derived.

II. We prove the existence of shadow Hamiltonians of any even order for arbitrary non-separable Hamiltonians and reversible symplectic integrators.

III. We introduce the shadow manifold Hamiltonian Monte Carlo algorithm (SMHMC), which is guaranteed to leave the target density invariant.

IV. Numerical examples are provided demonstrating advantages over existing alternatives.

V. A PyTorch implementation is made available at `https://github.com/chrisvdh/shadowtorch`.

2 Hamiltonian Monte Carlo

Markov chain Monte Carlo (MCMC) sampling techniques are a mainstay of computational statistics and machine learning. Many problems in scientific, medical, and industrial applications can be framed as Bayesian hierarchical problems, and their resolution often reduces to sampling from a posterior distribution in order to approximate some (often high-dimensional) analytically intractable integral. HMC is a highly successful MCMC technique that is somewhat resilient to the curse of dimensionality, without compromising convergence guarantees (Livingstone et al., 2016).

The goal of MCMC is to generate a set of correlated samples $\{\theta_i\}_{i=1}^n$ whose empirical distribution converges, as $n \to \infty$, to a target measure with Lebesgue density π_θ on $\mathbb{R}^d$. This enables the approximation of the integral $\mathbb{E}_\pi f = \int_{\mathbb{R}^d} f(\theta)\pi_\theta(\theta)\,d\theta$ of a π_θ-integrable function $f : \mathbb{R}^d \to \mathbb{R}$ via the Monte Carlo estimate $n^{-1}\sum_{i=1}^n f(\theta_i)$. In the sequel, we assume that all relevant functions are infinitely differentiable, although this can be relaxed in practice.

2.1 The Basic Setup

For a target density π_θ, let $U : \mathbb{R}^d \to \mathbb{R}$ be a *potential energy* function on the configuration space of a Hamiltonian system such that $\pi_\theta(\theta) := e^{-U(\theta)}/\mathcal{Z}$, where $\mathcal{Z}$ is a constant that ensures that π_θ integrates to unity. Under this interpretation, "energy wells" with low potential energy correspond to regions of high probability. In the basic HMC setup, the configuration space is augmented with a conjugate momentum variable $p \in \mathbb{R}^d$, which is assigned a non-negative *kinetic energy* term $K : \mathbb{R}^d \to \mathbb{R}$, typically taken to be $K(p) = p^\top \Sigma^{-1} p/2$. Here, Σ is a real-valued positive-definite $d \times d$ matrix, commonly taken to be the identity or some other diagonal matrix. Under this choice, the conjugate momenta are assumed to be distributed via a multivariate Gaussian density $\pi_p(p) := e^{-K(p)}/\sqrt{(2\pi)^d|\Sigma|}$, where we denote by $|\Sigma|$ the determinant of Σ. The Hamiltonian of the system is then defined by $H(\theta, p) := U(\theta) + K(p)$,

with the joint density of θ and p satisfying

$$\pi_H(\theta, p) \coloneqq \frac{e^{-H(\theta,p)}}{\mathcal{Z}\sqrt{(2\pi)^d|\Sigma|}} = \pi_\theta(\theta)\pi_p(p). \quad (1)$$

RMHMC is a method introduced by Girolami & Calderhead (2011) that generalises HMC by endowing $\mathbb{R}^d$ with a Riemannian metric, and sampling from the resulting Riemannian manifold (M, G). In the basic implementation, the statistical manifold M is equipped with a metric G, taken to be a variant of the Fisher–Rao metric that incorporates the Bayesian prior. This metric gives a canonical way to measure the squared length of a momentum p at θ, which is incorporated into the kinetic energy term $K(\theta, p) = p^\top G^{-1}(\theta)p/2$. This term is now position-dependent, and the corresponding conditional density becomes

$$\pi_{p|\theta}(\theta, p) \coloneqq \frac{e^{-K(\theta,p)}}{\sqrt{(2\pi)^d|G(\theta)|}}.$$

The determinant term is then incorporated into the Hamiltonian

$$H(\theta, p) \coloneqq U(\theta) + \frac{1}{2}\log\left((2\pi)^d|G(\theta)|\right) + K(\theta, p), \quad (2)$$

and by construction the joint density is then given by

$$\pi_H(\theta, p) \coloneqq \frac{e^{-H(\theta,p)}}{\mathcal{Z}} = \pi_\theta(\theta)\pi_{p|\theta}(\theta, p).$$

Given some initial pair $z_0 = (\theta_0, p_0)$, (RM)HMC algorithms then draw a candidate sample from the joint density π_H by evolving the *phase* z_0 along a trajectory in phase space defined by Hamiltonian dynamics.

2.2 Hamiltonian Dynamics

The key component of HMC is *Hamilton's equations of motion*: for time $t \in \mathbb{R}$ and $z(t) = (\theta(t), p(t))$, let

$$\frac{d\theta^i}{dt} = \frac{\partial H}{\partial p^i}, \quad \frac{dp^i}{dt} = -\frac{\partial H}{\partial \theta^i}, \quad (3)$$

or more succinctly

$$\frac{dz}{dt} = J\nabla H(z), \quad \text{where} \quad J \coloneqq \begin{pmatrix} 0 & I_{d\times d} \\ -I_{d\times d} & 0 \end{pmatrix}, \quad (4)$$

and $I_{d\times d}$ is the d-dimensional identity matrix. The following three properties of Hamiltonian dynamics are key (see Neal (2010)).

1. Conservation of energy: It is immediate from equation (3) that the vector field dz/dt lies orthogonal to the gradient vector field of H; that is,

$$\frac{dH}{dt} = \sum_{i=1}^{d} \frac{\partial H}{\partial \theta^i}\frac{d\theta^i}{dt} + \frac{dp^i}{dt}\frac{\partial H}{\partial p^i} = (\nabla H)^\top J\nabla H = 0.$$

2. Reversibility: Let $\varphi_t^H(z_0)$ denote the unique solution at time t to (3) with initial value z_0. Since H is time-homogeneous, there holds

$$\varphi_t^H \circ \varphi_s^H(z_0) = \varphi_{s+t}^H(z_0); \quad \text{i.e.,} \quad \varphi_{-t}^H \circ \varphi_t^H(z_0) = z_0, \quad (5)$$

or, in other words, $\varphi_{-t}^H = (\varphi_t^H)^{-1}$. Furthermore, if K is an even function, then we can obtain the inverse mapping by negating the sign of p, applying φ_t, and then negating p's sign again.

3. Volume preservation: Taking an open neighbourhood $N \subset \mathbb{R}^{2d}$ and evolving N by φ_t reveals $m(\phi_t(N)) = m(N)$, where m is the Lebesgue measure on $\mathbb{R}^{2d}$. This is most easily seen by noting that the Hamiltonian vector field $J\nabla H$ is divergence free, that is,

$$\nabla \cdot J\nabla H = \sum_{i=1}^{d} \frac{\partial}{\partial \theta^i}\frac{\partial H}{\partial p^i} - \frac{\partial}{\partial p^i}\frac{\partial H}{\partial \theta^i} = 0.$$

Given a candidate sample obtained by evolving the system (3) up to time T, these three properties allow for a simple Metropolis–Hastings acceptance criterion. Reversibility of the dynamics implies reversibility of the resultant Markov chain — therefore, the sampler leaves the target density invariant. Conservation of energy suggests a means of detecting numerically approximated trajectories that have not been faithful to (3) by computing $H\big(z(T)\big) - H(z_0)$, and volume preservation removes the need to keep track of a Jacobian in this change of variables. This allows proposed samples to be accepted with probability

$$\alpha = \min\left\{1, \exp\left(H(\theta, p) - H(\theta(T), p(T))\right)\right\}, \quad (6)$$

and otherwise taking $(\theta, -p)$. This final momentum flip is to guarantee reversibility.

Finally, it is worth mentioning that while properties 1–3 are certainly desirable, they are by no means necessary for an effective sampling algorithm. Accurate samplers have been constructed that break each of these conditions (Lan et al., 2015; Radivojević & Akhmatskaya, 2019). In these scenarios, greater care must be taken to ensure invariance to the correct target density.

2.3 Symplectic Integrators

Exact solutions to (3) are accepted with probability one, but are unavailable in non-trivial settings. Instead, it is necessary to approximate the solutions via numerical integration techniques. Doing so no longer preserves the Hamiltonian, and so changes the acceptance probability. Approximation errors and their corresponding Metropolis–Hastings acceptance probabilities for samples drawn from a Bayesian logistic regression posterior

density on a standard dataset are shown in Figure 1. Fortunately, there is a well studied, highly accurate family of geometric numerical integrators that preserve volume. These are known as *symplectic* numerical integrators, whose one-step map $\Phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ satisfies $D\Phi(z)^\top J D\Phi(z) = J$, where $\big(D\Phi(z)\big)^\top$ denotes the $2d \times 2d$ Jacobian of Φ at z (Bou-Rabee & Sanz-Serna, 2017).

It is clear from this definition that the composition of two symplectic maps is also symplectic, which allows us to approximate dynamics along lengthy trajectories by composing chains of short symplectic steps. We will see in Section 3.1 that by enforcing our numerical integrator to be symplectic, we can compute a family of *modified* or *shadow Hamiltonians* that are conserved up to arbitrary order in the step size.

In the Euclidean case (HMC), the Hamiltonian is *separable*; that is, $H(\theta, p) = U(\theta) + K(p)$. Systems of this type admit a certain type of decomposition or *splitting*, since the Hamiltonian vector field also decomposes into $J\nabla H(z) = J\nabla U(\theta) + J\nabla K(p)$. This induces the pair of *split systems*

$$\frac{d\theta^i}{dt} = \frac{\partial K}{\partial p^i},\ \frac{dp^i}{dt} = 0 \ \text{ and } \ \frac{dp^i}{dt} = \frac{\partial U}{\partial \theta^i},\ \frac{d\theta^i}{dt} = 0. \quad (7)$$

Individually, these systems are analytically solvable for arbitrary times. However, doing so results in large errors, due to the effective decoupling caused by the splitting. By using a relatively small step size, we can mitigate this error by iterative updates, each of which exactly solves the split systems (7)

$$\begin{cases} \psi_h^U(\theta_{n+1}, p_n) := p_n - h\frac{\partial U}{\partial \theta}(\theta_{n+1}) = p_{n+1}, \\ \psi_h^K(\theta_n, p_n) := \theta_n + h\frac{\partial K}{\partial p}(p_n) = \theta_{n+1}. \end{cases} \quad (8)$$

While each of these steps is reversible, their composition is not. Instead, by composing (8) with its reversal, we obtain the palindromic *Störmer-Verlet* scheme

$$\begin{cases} p_{n+\frac{1}{2}} = p_n - \frac{h}{2}\frac{\partial U}{\partial \theta}(\theta_n), \\ \theta_{n+1} = \theta_n + h\frac{\partial K}{\partial p}(p_{n+\frac{1}{2}}), \\ p_{n+1} = p_{n+\frac{1}{2}} - \frac{h}{2}\frac{\partial U}{\partial p}(\theta_{n+1}), \end{cases} \quad (9)$$

which is accurate up to $\mathcal{O}(h^2)$. This integrator is reversible, since $(\psi^U_{-h/2} \circ \psi^K_{-h} \circ \psi^U_{-h/2}) \circ (\psi^U_{h/2} \circ \psi^K_h \circ \psi^U_{h/2})(z_0) = z_0$. A number of higher-order integrators have also been studied (Yoshida, 1990; Radivojević et al., 2018), and more recently applied to problems in Bayesian statistics (Radivojević & Akhmatskaya, 2019).

In order to approximate a trajectory up to time $T = hL$ for some $L \in \mathbb{N}$, we simply compose (9) L times

$$\Phi_h^N := \overbrace{(\psi^{\hat{U}}_{h/2} \circ \psi^{\hat{K}}_h \circ \psi^{\hat{U}}_{h/2}) \circ \cdots \circ (\psi^{\hat{U}}_{h/2} \circ \psi^{\hat{K}}_h \circ \psi^{\hat{U}}_{h/2})}^{L \text{ times}}. \quad (10)$$

When considering Hamiltonians of the form (2), updates of the form (8) based on approximations given by (7) do not have unit Jacobian determinant, and do not provide reversible updates (Girolami & Calderhead, 2011). Instead, standard implementations employ an *implicit* numerical integrator called the *generalized leapfrog algorithm* to address these issues.

The generalized leapfrog method is constructed in the same way as the leapfrog method. By composing two partitioned Euler methods, a symplectic method is obtained (Hairer et al., 2006, VI.3). The updates with step-size h are of the form

$$\begin{cases} p_{n+\frac{1}{2}} = p_n - \frac{h}{2}\nabla_\theta H\big(\theta_n, p_{n+\frac{1}{2}}\big), \\ \theta_{n+1} = \theta_n + \frac{h}{2}\nabla_p\Big[H\big(\theta_n, p_{n+\frac{1}{2}}\big) + H\big(\theta_{n+1}, p_{n+\frac{1}{2}}\big)\Big], \\ p_{n+1} = p_{n+\frac{1}{2}} - \frac{h}{2}\nabla_\theta H\big(\theta_{n+1}, p_{n+\frac{1}{2}}\big). \end{cases} \quad (11)$$

This method is again accurate up to second order in h, and again, we compute the trajectory up to time $T = hL$ by composing (11) L times, analogous to (10). The distinguishing feature here is that each variable on the left hand side of the first two lines in (11) also appear on the right hand side, that is, it is *implicitly defined*.

The immediate drawback of implicit numerical methods is that, in contrast to (8), the sub-problems can no longer be exactly solved. Fixed point iterations are commonly used to approximate the updates, although other implicit methods for Euclidean HMC have been proposed that employ Newton's method (Pourzanjani & Petzold, 2019). Recently, an explicit method has also been developed (Cobb et al., 2019) based on an approach for approximating the dynamics of systems with non-separable Hamiltonians in the physics literature (Tao, 2016; Pihajoki, 2015). This technique does so by simulating parallel problems on an extended phase space (the product space of two copies of the cotangent bundle), with a 'binding step' in between. The integrator is computationally cheaper than (11), and has a known fourth-order shadow Hamiltonian (Tao, 2016). However, it is not clear how to incorporate this into a shadow HMC algorithm in a way that ensures the resulting sampler leaves the target density invariant. Furthermore, it introduces an extra binding hyperparameter, which the stability of the trajectories appear to depend upon in a delicate way.

3 Shadow Manifold Hamiltonian Monte Carlo

HMC methods rely on the numerical integrators' accurate simulation of the trajectories dictated by Hamiltonian dynamics. However, leapfrog and generalized leapfrog integrators only preserve the Hamiltonian up to second order. In order to increase accuracy, one could design more accurate integrators, although these tend to be computationally expensive. Our approach relies on backwards error analysis to instead derive a *shadow Hamiltonian*, whose energy is more accurately conserved by the generalized leapfrog algorithm. By instead targeting the corresponding *shadow density*, we see higher sample acceptance probabilities (Figure 1). Importance sampling is then employed to correct these samples towards the true density. This efficient combination of MCMC and importance sampling allows for even greater performance than either strategy alone.

3.1 Shadow Hamiltonians

Shadow Hamiltonians are commonly obtained by truncating the *Baker–Campbell–Hausdorff* (BCH) formula applied to Poisson brackets of the terms of a separable Hamiltonian (Barp et al., 2018). Since short-time solutions to Hamilton's equations are realized as an exponential of the Hamiltonian vector field, the BCH formula gives us a way to track the approximation error induced by iterative composition of the non-commutative split vector fields. The underlying algebraic relationship between Hamiltonian vector fields and Hamiltonians provides a clean representation in terms of an asymptotic expansion of nested Poisson brackets. This expansion, when truncated to an appropriate order, reveals the shadow Hamiltonian.

Recall that for the leapfrog integrator (8), the Hamiltonian vector field can be split into two orthogonal vector fields, $J\nabla U$ and $J\nabla K$. By applying the BCH formula to each composition in (9), the fourth–order shadow Hamiltonian for the leapfrog integrator with step size h (Radivojević & Akhmatskaya, 2019) is obtained:

$$\begin{aligned}\mathcal{H}_L &= H + \frac{h^2}{12}\{K,\{K,U\}\} - \frac{h^2}{24}\{U,\{U,K\}\} \qquad (12)\\ &= H + \frac{h^2}{12}p^\top M^{-1}U_{\theta\theta}M^{-1}p - \frac{h^2}{24}\nabla U^\top M^{-1}\nabla U,\end{aligned}$$

where the Poisson bracket satisfies $\{f,g\} = \frac{\partial f}{\partial \theta^i}\frac{\partial g}{\partial p^i} - \frac{\partial f}{\partial p^i}\frac{\partial g}{\partial \theta^i}$. Here, we denote the Hessian of U by $U_{\theta\theta}$ and suppress dependence on θ and p.

However, the generalized leapfrog integrator is more delicate. The BCH formula can only be naively applied to time-homogeneous systems of ODEs, and its implicit nature means that it can no longer be split into time-homogeneous Hamiltonian vector fields. Taking a continuous extension of (11) in t, we have that $(\theta(t), p(t))$ satisfies

$$\begin{cases} q(t) = p_0 - \frac{t}{2}\nabla_\theta H\big(\theta(t), q(t)\big), \\ \theta(t) = \theta_0 + \frac{t}{2}\nabla_p\Big[H\big(\theta_0, q(t)\big) + H\big(\theta(t), q(t)\big)\Big], \\ p(t) = q(t) - \frac{t}{2}\nabla_\theta H\big(\theta(t), q(t)\big). \end{cases} \qquad (13)$$

Differentiating and solving for $(d\theta/dt, dp/dt)$, we find that the vector field governing the system varies with time. This necessitates a different approach.

To compute the fourth–order shadow Hamiltonian, we must first identify its corresponding vector field up to third order. This is done by a comparing the third–order Taylor expansions of the discretized solution $\big(\hat\theta(h), \hat p(h)\big)$, and the analytic solution $\big(\theta(h), p(h)\big)$. By appropriately integrating the vector field with respect to θ and p, we can then identify the fourth–order shadow Hamiltonian $\mathcal{H}^{[4]}$.

Theorem 1. *Let $M = \mathbb{R}^d$, and $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a smooth Hamiltonian function. The fourth–order shadow Hamiltonian function $\mathcal{H}^{[4]} : \mathbb{R}^d \to \mathbb{R}$ corresponding to the generalized leapfrog integrator is given by*

$$\begin{aligned}\mathcal{H}^{[4]}(\theta,p) = H(\theta,p) + \frac{h^2}{12}\Big(&\nabla_p H\nabla_{\theta\theta}H\nabla_p H \qquad (14)\\ &- \frac{1}{2}\nabla_\theta H\nabla_{pp}H\nabla_\theta H\\ &+ \nabla_\theta H\nabla_{\theta p}H\nabla_p H\Big).\end{aligned}$$

Comparing equations (14) and (12), we note an additional term containing mixed partial derivatives of the Hamiltonian in θ and p. Consequently, when the Hamiltonian is separable (as in the Euclidean case), we recover (12).

The complete proof is tedious, and is relegated to the supplementary material. In our implementation, the second term in (14) is computed using quantities already required for the dynamics, while the other terms of the shadow Hamiltonian can be computed using automatic differentiation. Judicious use of matrix-vector products allows this to be done in a way that doesn't dominate the $\mathcal{O}(d^3)$ complexity of matrix inversion or decomposition. Difference quotient approximations, such as those used by Radivojević & Akhmatskaya (2019), could also be applied.

Proceeding further, this construction can be extended to obtain shadow Hamiltonians of arbitrary even order. Since modern statistical applications are increasingly concerned with more exotic spaces, it is natural to ask if such a construction could be considered on, for example, spheres and Stiefel manifolds. Hairer (2003)

provides local results, and a construction when the manifold M is given by a level set of a known smooth function. For another broad class of manifolds, and any reversible symplectic numerical integrator, we have the following global result.

Theorem 2. *Let M be a smooth simply connected Riemannian manifold with cotangent bundle T^*M, and let $H : T^*M \to \mathbb{R}$ be smooth Hamiltonian function. Then for any fixed reversible symplectic integrator, there exists a family of shadow Hamiltonians $\mathcal{H}^{[2k]} : T^*M \to \mathbb{R}$ indexed by $k \in \mathbb{N}$, such that $\mathcal{H}^{[2k]}$ is preserved by the integrator* (11) *with step size h up to $\mathcal{O}(h^{2k})$.*

Here we have required the topological property that the base manifold is simply connected. Roughly speaking, this means that the manifold has no 'holes', which is satisfied by a wide class of manifolds that may be of interest in statistical applications. However, known HMC methods on these classes of manifolds are either constructed using constrained Hamiltonians (Brubaker et al., 2012), or directly exploit additional structure in their geometry (Barp et al., 2019). In either case, the generalized leapfrog integrator is not used, so Theorem 1 cannot be used directly. However, the procedure used in its proof can be applied to obtain a local shadow Hamiltonian, providing avenues for future research.

The proof of Theorem 2, also relegated to supplementary material, is a straightforward consequence of de Rham's theorem (Warner, 2013). While this is a basic result in algebraic and differential topology, it's largely unknown to the statistics and machine learning communities. We remark that for fixed positive h, one may expect $\mathcal{H}^{[2k]}$ to diverge as $k \to \infty$.

Upon inspection of (14), we notice that Hamiltonians of the form (2) may not amount to desirable tail behaviour, which can cause difficulties in integrability of the resulting shadow density, and may provide an obstruction to geometric ergodicity. To remedy this, we can follow Izaguirre & Hampton (2004) and set

$$\mathcal{H}(\theta, p) := \max\left\{\mathcal{H}^{[4]}(\theta, p) + c, H(\theta, p)\right\}, \qquad (15)$$

where c is a constant that is chosen by the practitioner. This guarantees that the tail behaviour of $\mathcal{H}$ is always dictated by the Hamiltonian itself, which in turn controls the asymptotic decay of the density and the performance of subsequent importance sampling.

3.2 Momentum Refreshment

A key distinction between shadow Hamiltonian methods and regular HMC is that the shadow density's conditional density $\pi_{\mathcal{H}}(p \,|\, \theta)$ is no longer Gaussian. This means that naïvely sampling new Gaussian momenta and accepting with probability one no longer leaves the conditional density invariant. Moreover, we can no longer expect to be able to analytically marginalize out the momenta. This was initially circumvented via rejection sampling (Izaguirre & Hampton, 2004). At each refreshment step a new momentum proposal can be generated $\hat{p} \sim \mathcal{N}(0, G(\theta))$, and accepted with probability $\min\left\{1, \exp(\mathcal{H}(\theta, \hat{p}) - \mathcal{H}(\theta, p))\right\}$. If the new momentum is rejected, this is repeated until a sample is accepted. This results in a method that more accurately reflects the original implementation of HMC; however, repeated rejection can result in many computationally expensive shadow Hamiltonian evaluations.

More recent approaches utilize partial momentum refreshment via an additional Metropolis–Hastings step; see Kennedy & Pendleton (2001); Akhmatskaya & Reich (2006) and also Radivojević & Akhmatskaya (2019) for a brief review. In this regime, an auxiliary noise vector $u \sim \mathcal{N}(0, G(\theta))$ is drawn and a momentum proposal is generated via the mapping $R : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ that satisfies

$$R(p, u) = (\rho\, p + \sqrt{1 - \rho^2}\, u, -\sqrt{1 - \rho^2}\, p + \rho\, u).$$

The new parameter $\rho = \rho(\theta, p, u)$ takes values in $(0, 1]$ and controls the extent of the momentum retention. The proposals are then accepted according to the modified density corresponding to $\overline{\mathcal{H}}(\theta, p, u) = \mathcal{H}(\theta, p) + \frac{1}{2} u^\top G^{-1}(\theta) u$. The updated momentum is then taken to be

$$\begin{cases} \rho\, p + \sqrt{1 - \rho^2}\, u & \text{with probability } \gamma, \\ p & \text{otherwise,} \end{cases}$$

where

$$\gamma := \max\left\{1, \exp\left(\overline{\mathcal{H}}(\theta, p, u) - \overline{\mathcal{H}}(\theta, R(p, u))\right)\right\}. \quad (16)$$

This procedure results in a Markov chain that preserves some of the dynamics between subsequent samples. Note that ρ introduces another degree of freedom into the sampler, and can be chosen to depend on θ and p.

3.3 The SMHMC Sampler

An algorithmic description of the SMHMC sampler is provided in Algorithm 1. Since the sampler now involves a composition of two reversible Metropolis–Hastings steps, the resulting Markov chain is no longer reversible. By breaking the detailed balance relations, it is no longer immediately clear that the target density is stationary, and so this must be demonstrated. To this end we have the following guarantee, the proof of which is taken from Radivojević & Akhmatskaya (2019) or Fang et al. (2014), after noting that the explicit form of the shadow density plays no role.

Theorem 3. *The SMHMC sampler leaves the target density $\pi_{\mathcal{H}}$ invariant.*

Algorithm 1: SMHMC

Input: maximum number of steps L, step size h, Riemannian metric G, momentum update parameter ρ, number of Monte Carlo samples n, initial values (θ_0, p_0).
for $i = 1$ **to** k **do**
  sample number of steps l from $\{1, \dots, L\}$.
  store $(\theta, p) \leftarrow (\theta_{i-1}, p_{i-1})$.
  sample momentum update proposal $u \sim \mathcal{N}(0, G(\theta))$.
  update $\overline{p} \leftarrow \rho\, p + \sqrt{1-\rho^2}\, u$ with probability γ, defined in (16).
  compute the Shadow Hamiltonian $\mathcal{H}(\theta, \overline{p})$.
  integrate Hamiltonian Dynamics $(\hat{\theta}, \hat{p}) \leftarrow \Phi_h^l(\theta, \overline{p})$.
  accept sample $(\theta_i, p_i) \leftarrow (\hat{\theta}, \hat{p})$ with probability β, and reject $(\theta_i, p_i) \leftarrow (\theta, -p)$ otherwise. Here, $\beta = \min\{1, \exp(-\Delta\mathcal{H})\}$.
  compute the importance sampling weights $w_i = \exp\big(\mathcal{H}(\theta_i, p_i) - H(\theta_i, p_i)\big)$
end for
return $(\theta_i, p_i, w_i)_{i=0}^k$

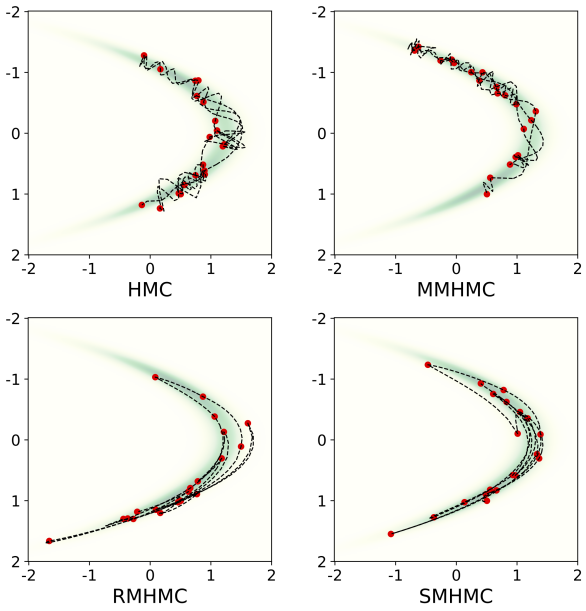

Figure 2: Trajectories of the first 20 samples of the banana-shaped density drawn by HMC, MMHMC, RMHMC and SMHMC. Accepted samples are denoted by red dots.

Of course, the target density of the SMHMC sampler is proportional to $\exp\big(-\mathcal{H}(\theta, p)\big)$, and is therefore not the target density π_θ in general. In order to correct for this bias, samples can then be reweighted via importance sampling, with the i-th sample's weight given by the *exponential shadows* $w_i = \exp\big(\mathcal{H}(\theta_i, p_i) - H(\theta_i, p_i)\big)$. Integrals of the form $\int_{\mathbb{R}^d} f(\theta)\pi_\theta(\theta)\, d\theta = \int_{\mathbb{R}^d \times \mathbb{R}^d} f(\theta)\frac{\pi_H(\theta,p)}{\pi_{\mathcal{H}}(\theta,p)}\pi_{\mathcal{H}}(\theta, p)\, d\theta\, dp$ can then be approximated via the importance sampling estimator $\mathrm{I}_n(f) = \sum_{i=1}^n \bar{w}_i f(\theta_i)$, where $\bar{w}_i = w_i/(\sum_{i=1}^n w_i)$ are the normalized weights.

One potential drawback of importance sampling is its deterioration in performance as the sampling and target densities grow far apart. Fortunately, the smoothness of (14) guarantees pointwise control over the behaviour of the weights in the continuous limit. A Taylor expansion of w_i in view of (14) shows that $w_i = 1 + \mathcal{O}(h^2)$ almost surely as $h \to 0^+$. The performance in the importance sampling step should be adequate provided h is not too large.

4 Numerical Experiments

We consider three test problems to demonstrate the dynamics and performance of SMHMC. The first is a toy example that illustrates the dynamics of SMHMC compared to HMC, MMHMC (Radivojević & Akhmatskaya, 2019) and RMHMC. We then consider performance in a basic statistical problem, namely a Bayesian Logistic Regression model, over various benchmark datasets. This is a standard example where RMHMC outperforms HMC in low dimensions, but begins to struggle in higher dimensions. We then turn our attention to Neal's funnel problem (Neal, 2003), which is a simple example that captures the pathological geometric features typical of those found in Bayesian hierarchical models. Performance is measured via sample acceptance rate, minimum effective sample size (ESS), and ESS per second; details on their computation are provided in the supplementary material.

4.1 Banana-shaped Density

The banana-shaped density is a variant of the Rosenbrock function suggested by Bornn and Cornebise in their discussion of Girolami & Calderhead (2011) as an example with strong ridge-like geometric features typical of those found in non-identifiable models. We tested the dynamics of SMHMC, RMHMC, HMC and MMHMC in sampling the two-dimensional posterior density $\pi(\theta \,|\, y)$ based on the following model:

$$y \,|\, \theta \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2), \quad \theta_1, \theta_2 \sim \mathcal{N}(0, \sigma_\theta^2).$$

Following Lan et al. (2015); Radivojević & Akhmatskaya (2019), one hundred data points $\{y_i\}_{i=1}^{100}$ are generated with $\theta_1 + \theta_2^2 = 1$, $\sigma_y = 2$ and $\sigma_\theta = 1$. Samples are then drawn from the posterior density across our range of samplers. The dynamics of the HMC, MMHMC, RMHMC and SMHMC samplers

Table 1: Summary of acceptance rates, minimum effective sample size, and minimum effective sample size per second for posterior samples of Bayesian logistic regression on four datasets. Reported values are averaged across 10 chains.

				Acceptance		minESS		minESS/s	
Data set	d	h	α	RMHMC	SMHMC	RMHMC	SMHMC	RMHMC	SMHMC
Australian	15	0.5	100	0.9237	**0.9929**	5167.19	**6212.87**	**56.0732**	52.6366
German	25	0.5	100	0.8933	**0.9727**	3758.42	**5452.45**	35.1639	**40.9878**
Parkinson's	46	0.3	1	0.9391	**0.9933**	1553.08	**2469.34**	14.4799	**17.8577**
Sonar	61	0.3	1	0.8898	**0.9639**	1371.66	**2273.69**	9.5294	**12.9841**

are illustrated in Figure 2. This example clearly delineates the effect of the geometric adaptation on the sampler's dynamics, while the partial momentum retention's effects are less pronounced.

4.2 Bayesian Logistic Regression

Bayesian Logistic Regression is a standard tool for binary classification that is ubiquitous in the medical, scientific, engineering, and financial fields to name a few (Gelman et al., 2004). We present results from the analysis of four datasets, retrieved from the UCI Machine Learning Repository (Dua & Graff, 2017).

As a common practice, the dataset is centred and normalized to have zero mean and unit variance in each dimension. The potential $U(\theta)$ is then taken to be the negative log–likelihood of the data at θ, and a Gaussian prior with variance α on each of the parameters is imposed. We took the metric G to be the prior-adapted Fisher information from Girolami & Calderhead (2011), estimated via the (positive-definite) negative empirical Hessian. To this end, 10 chains of 5000 samples were drawn from the posterior of each dataset, after 500 samples of burn–in. We set $\rho = 0.25$ in the SMHMC sampler. Results are presented in Table 1.

4.3 Neal's Funnel Density

The funnel density was suggested by Neal (2003) as a sampling problem that exhibits behaviour typical of pathologies that arise in Bayesian hierarchical and latent variable models, particularly those with sparse datasets (Betancourt & Girolami, 2015). The model treats the variance of the parameters as a latent log-normal random variable, which leads to non-convex exponential cusping behaviour in the negative log-density. Neal (2003) considered a 10-dimensional funnel — here, we instead consider a 30-dimensional example. The target density is defined to satisfy $\pi(\theta, v) := \mathcal{N}\big(v \,|\, 0, 9\big) \prod_{i=1}^{29} \mathcal{N}\big(\theta_i \,|\, 0, e^v\big)$. This is a problem for which HMC typically demonstrates poor performance (Betancourt, 2013; Cobb et al., 2019), and struggles to draw samples from inside the 'throat' of the funnel. Since the negative log-density is highly non-convex, the SoftAbs metric is employed to give a positive definite approximation of the expected Hessian, while retaining its eigenvectors (Betancourt, 2013).

We ran 10 chains of 2000 samples each for the HMC, RMHMC and SMHMC samplers on a 30-dimensional funnel. Following a few pilot runs, a step-size of $h = 0.3$, with a maximum of 64 leapfrog steps per sample trajectory was chosen for the geometric algorithms. For HMC, these hyperparameters were set to 0.075 and 500.

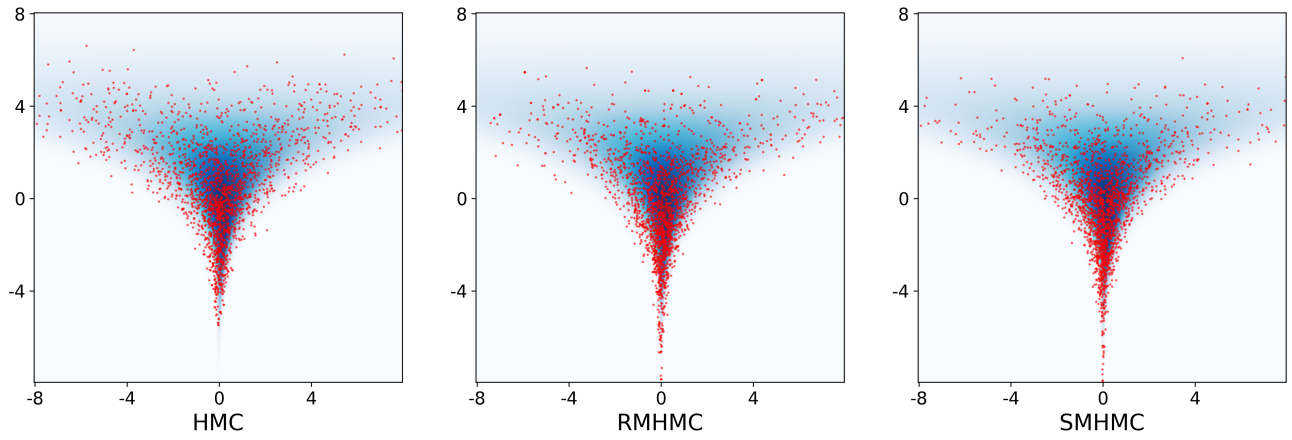

Figure 3: Samples drawn via HMC, RMHMC and SMHMC on the first two dimensions of a 30-dimensional funnel.

Table 2: Summary of acceptance rates, minimum effective sample size, and minimum effective sample size per second for samples drawn from funnel density. Reported values are averaged across 10 chains.

	ACCEPT	MINESS	MINESS/S
HMC	0.9771	7.32	0.0220
RMHMC	0.9576	397.48	1.0311
RMHMCρ	0.9640	550.01	1.3706
SMHMC	**0.9878**	545.29	1.3167
SMHMCρ	0.9843	**578.25**	**1.4003**

The momentum retention ρ was set to zero and 0.25 in the RMHMC/SMHMC and RMHMCρ/SMHMCρ variants, respectively. Figure 3 shows the locations of these samples. The SMHMC sampler exhibits comparable throat penetration to the RMHMC sampler. Acceptance rates, minimum ESS and minimum ESS per second are reported in Table 2.

5 Conclusion

We have extended shadow Hamiltonian Monte Carlo methods by deriving the shadow Hamiltonian generated by the generalized leapfrog numerical integrator. The resulting SMHMC algorithm shows promise as a means of pushing past the limitations of RMHMC while retaining the advantages of adapting the sampler to the geometry of the target density. SMHMC shows promise for deployment in complex higher-dimensional hierarchical Bayesian models, where the dimensionality causes RMHMC's performance to wane.

Acknowledgements

This work has been supported by the Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS), under grant number CE140100049.

References

Akhmatskaya, E. and Reich, P. S. The Targeted Shadowing Hybrid Monte Carlo (TSHMC) Method. In *New Algorithms for Macromolecular Simulation. Lecture Notes in Computational Science and Engineering, vol 49*, pp. 1–51. Springer, 2006.

Amari, S.-I. *Information Geometry and Its Applications*. Springer, Berlin, 1st edition, 2016.

Barp, A., Briol, F.-X., Kennedy, A. D., and Girolami, M. Geometry and Dynamics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 5(1):451–471, 2018.

Barp, A., Kennedy, A., and Girolami, M. Hamiltonian Monte Carlo on Symmetric and Homogeneous Spaces via Symplectic Reduction. *arXiv preprint arXiv:1903.02699*, 2019.

Betancourt, M. A general metric for Riemannian manifold Hamiltonian Monte Carlo. In Nielsen, F. and Barbaresco, F. (eds.), *In First International Conference on the Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pp. 327–334, 2013.

Betancourt, M. and Girolami, M. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 54:79—-101, 2015.

Betancourt, M., Byrne, S., Livingstone, S., and Girolami, M. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23, 2017.

Bou-Rabee, N. and Sanz-Serna, J. M. Yet more elementary proofs that the determinant of a symplectic matrix is 1. *Linear Algebra and its Applications*, 515, 2017.

Brubaker, M., Salzmann, M., and Urtasun, R. A family of MCMC methods on implicitly defined manifolds. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pp. 161–172, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Cobb, A., Baydin, A., Markham, A., and Roberts, S. Introducing an explicit symplectic integration scheme for Riemannian manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:1910.06243*, 2019.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics Letters B*, 1987.

Fang, Y., Sanz-Serna, J., and Skeel, R. Compressible Generalized Hybrid Monte Carlo. *The Journal of Chemical Physics*, 140:174108, 05 2014.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Hairer, E. Global modified Hamiltonian for constrained symplectic integrators. *Numerische Mathematik*, 95 (2):325–336, 2003.

Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration*. Springer, Berlin, 2nd edition, 2006.

Hatcher, A. *Algebraic topology*. Cambridge Univ. Press, Cambridge, 2000.

Horowitz, A. M. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 1991.

Izaguirre, J. and Hampton, S. Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules. *Journal of Computational Physics*, 200, 2004.

Kennedy, A. and Pendleton, B. Cost of the generalised hybrid Monte Carlo algorithm for free field theory. *Nuclear Physics B*, 2001.

Lan, S., Stathopoulos, V., Shahbaba, B., and Girolami, M. Markov Chain Monte Carlo from Lagrangian Dynamics. *Journal of Computational and Graphical Statistics*, 24, 2015.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. On the Geometric Ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25:3109–3138, 2016.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer, Berlin, 1996.

Neal, R. M. Slice sampling. *Ann. Statist.*, 31(3):705–767, 06 2003.

Neal, R. M. MCMC using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

Pihajoki, P. Explicit methods in extended phase space for inseparable hamiltonian problems. *Celestial Mechanics and Dynamical Astronomy*, 121(3), 2015.

Pourzanjani, A. and Petzold, L. Implicit Hamiltonian Monte Carlo for Sampling Multiscale Distributions. *arXiv preprint arXiv:1911.05754*, 11 2019.

Radivojević, T. and Akhmatskaya, E. Modified Hamiltonian Monte Carlo for Bayesian inference. *Statistics and Computing*, 2019.

Radivojević, T., Fernández-Pendás, M., Sanz-Serna, J. M., and Akhmatskaya, E. Multi-stage splitting integrators for sampling with modified Hamiltonian Monte Carlo methods. *Journal of Computational Physics*, 373, 2018.

Sohl-Dickstein, J. Hamiltonian Monte Carlo with Reduced Momentum Flips. *arXiv preprint arXiv:1205.1939*, 2012.

Sohl-Dickstein, J., Mudigonda, M., and Deweese, M. Hamiltonian Monte Carlo Without Detailed Balance. *Proceedings of the 31st International Conference on Machine Learning*, 32, 2014.

Tao, M. Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance. *Phys. Rev. E*, 94, 2016.

Warner, F. W. *Foundations of differentiable manifolds and Lie groups; 2nd ed.* Graduate texts in mathematicals. Springer, New York, NY, 1983. doi: 10.1007/978-1-4757-1799-0.

Warner, F. W. *Foundations of differentiable manifolds and Lie groups*, volume 94. Springer Science & Business Media, 2013.

Whitehead, G. W. *Elements of homotopy theory*. Graduate Texts in Mathematics. Springer, New York, 1978. doi: 10.1007/978-1-4612-6318-0.

Yoshida, H. Construction of higher order symplectic integrators. *Physics Letters A*, 150(5), 1990.